Procédé de correspondance automatique entre des éléments graphiques et des éléments phonétiques

La présente invention concerne en général l'extraction automatique de connaissances linguistiques dans un corpus de transcriptions de chaînes graphiques en des chaînes phonétiques. Plus particulièrement, elle concerne la transcription d'éléments typographiques tels que des caractères dans une langue prédéterminée en des éléments phonétiques.

Actuellement, chaque mot d'une langue constitue une chaîne graphique qui est transcrite phonétiquement en une chaîne de phonèmes par un phonéticien. Pour tout nouveau mot à ajouter à un corpus d'apprentissage, le phonéticien doit intervenir pour transcrire phonétiquement ce nouveau mot. Le corpus d'apprentissage ne fournit ainsi que des transcriptions graphème/phonème globales. Par exemple dans la transcription globale "ruelle"/[ryɛl], le corpus indique que globalement, la chaîne graphique "ruelle" se traduit en chaîne phonétique. Cependant, il n'est pas explicité que de quelque manière, unitairement, l'élément typographique "r" se retranscrit phonétiquement. La transcription globale n'indique pas également les syllabes ou graphèmes composant la chaîne graphique et les éléments phonétiques composant la chaîne phonétique.

Or la connaissance de la transcription élémentaire de chaque élément typographique permet, par la suite, par analyse caractère par caractère de toute chaîne graphique, de déterminer une ou plusieurs chaînes phonétiques associées à la chaîne graphique. Les transcriptions phonétiques sont utiles

à des systèmes correcteurs de fautes pour reconnaître des fautes lexicales lors de la saisie de texte sur un clavier. Il existe donc un besoin à partir d'une transcription brute d'extraire des transcriptions élémentaires plus fines.

L'invention vise à déduire automatiquement des transcriptions brutes de chaînes graphiques, telles que mots et noms patronymiques, par exemple, en des chaînes phonétiques, des transcriptions d'éléments graphiques, telles que caractères, en des éléments phonétiques composant les chaînes phonétiques afin de segmenter automatiquement toute chaîne graphique en graphèmes et toute chaîne phonétique en phonèmes. Les transcriptions élémentaires élément graphique par élément graphique, c'est-à-dire caractère par caractère, facilitent ensuite la transcription globale automatique de toute chaîne graphique supplémentaire apportée au corpus des chaînes graphiques, sur la base notamment d'une concaténation d'éléments phonétiques correspondant de manière biunivoque aux caractères de la chaîne graphique supplémentaire.

A cette fin, un procédé selon l'invention fait correspondre automatiquement des éléments graphiques composant des chaînes graphiques données à des éléments phonétiques composant des chaînes phonétiques correspondantes, après avoir saisi initialement des transcriptions globales des chaînes graphiques en les chaînes phonétiques dans une base accessible par l'ordinateur et avoir estimé et enregistré dans la base des premières probabilités de transcriptions élémentaires des éléments graphiques

respectivement en les éléments phonétiques. Le procédé est caractérisé par les étapes suivantes :

pour chaque transcription d'une chaîne graphique donnée à M éléments graphiques en une chaîne phonétique correspondante à N éléments phonétiques, déterminer des deuxièmes probabilités de MN deuxièmes transcriptions de M chaînes graphiques concaténant successivement les M éléments graphiques en N chaînes phonétiques concaténant successivement les N éléments phonétiques, en fonction chacune d'une première probabilité respective et de la plus grande de trois deuxièmes probabilités respectives déterminées précédemment, et

établir et mémoriser un lien entre les derniers éléments des chaînes graphique et phonétique de chaque deuxième transcription et les derniers éléments des chaînes graphique et phonétique de la transcription relative à la plus grande des trois deuxièmes probabilités respectives afin que des liens établis dans une matrice de taille MN relative aux deuxièmes probabilités constitue un chemin unique entre des dernier et premier couples d'éléments graphique et phonétique de la matrice pour segmenter la chaîne graphique donnée en des graphèmes correspondant respectivement à des phonèmes segmentant la chaîne phonétique correspondante et pour enregistrer les correspondances entre les graphèmes et phonèmes dans la base, le nombre d'éléments graphiques dans un graphème étant identique au nombre d'éléments phonétiques dans le phonème correspondant, afin que toute nouvelle chaîne graphique soit automatiquement transcrite en une chaîne phonétique segmentée en phonèmes au moyen des correspondances enregistrées.

Selon d'autres caractéristiques de l'invention, la première probabilité respective pour la détermination d'une deuxième probabilité relative à une deuxième transcription d'une chaîne graphique concaténant m éléments graphiques en une chaîne phonétique concaténant n éléments phonétiques, avec $1 \leq m \leq M$ et $1 \leq n \leq N$, est relative aux derniers éléments dans la chaîne graphique à m éléments graphiques et la chaîne phonétique à n éléments phonétiques. Les trois deuxièmes probabilités respectives déterminées précédemment pour la deuxième transcription de la chaîne graphique à m éléments graphiques en la chaîne phonétique à n éléments phonétiques sont de préférence respectivement relatives à une deuxième transcription d'une chaîne graphique à m-1 éléments graphiques en la chaîne phonétique à n éléments phonétiques, une deuxième transcription de la chaîne graphique à m éléments graphiques en une chaîne phonétique à n-1 éléments phonétiques et une deuxième transcription de la chaîne graphique à m-1 éléments graphiques en la chaîne phonétique à n-1 éléments phonétiques.

Par exemple, l'invention transcrit phonétiquement à partir du corpus de transcriptions globales telles que "ruelle"|[ryɛl] les éléments graphiques "r", "u", "e", "lle" respectivement en les éléments phonétiques [r], [y], [ɛ], [l].

L'invention peut être assimilée à une syllabation qui permet par analyse de décomposer une transcription globale en transcriptions élémentaires, et de mettre en correspondance localement des sous-transcriptions graphème/phonème. Le découpage en graphèmes et phonèmes initiaux et la mise en correspondance biunivoque de chaque élément graphique à chaque élément phonétique des phonèmes découpés est

appelée alignement graphème|phonème. Selon l'exemple précédent, l'invention produit l'alignement suivant :

"r" "u" "e" "lle"

[r] [y] [ɛ] [l**].

Le symbole * désigne un élément phonétique muet et sans signification.

D'autres caractéristiques et avantages de la présente invention apparaîtront plus clairement à la lecture de la description suivante de plusieurs réalisations préférées de l'invention, à titre d'exemples non limitatifs, en référence aux dessins annexés correspondants dans lesquels :

- la figure 1 est un algorithme d'étapes principales du procédé de correspondance automatique selon l'invention; et
- la figure 2 est un algorithme de sous-étapes d'une étape de détermination de premières probabilités individuelles incluse dans le procédé de correspondance automatique.

Comme montré à la figure 1, le procédé de correspondance automatique d'éléments graphiques et d'éléments phonétiques selon l'invention comprend des étapes principales E1 à E11. Ces étapes sont pour la plupart mises en oeuvre par exemple sous la forme d'un logiciel implémenté dans un terminal, tel qu'un ordinateur personnel ou un mobile dans un réseaun de radiocommunication cellulaire, et lié notamment à un système logiciel de correction orthographique de fautes lexicales qui peut être integré à un système de traitement de texte ou à un système d'exercice linguistique. Le terminal contient ou peut accéder à une base de données du type de celles utilisées en

intelligence artificielle. La base mémorise un corpus C de transcriptions globales initiales.

Initialement à l'étape E1, les transcriptions globales (CG|CP) sont constituées par des couples faisant correspondre chacun une chaîne graphique CG, telle qu'un mot dans une langue prédéterminée ou un nom patronymique, à une chaîne phonétique CP. Ces transcriptions ont été déterminées et saisies par un phonéticien au moyen d'un formulaire adéquat affiché par l'ordinateur. Le corpus C fait correspondre des chaînes graphiques GC composées chacune d'un ou plusieurs éléments typographiques (caractères), appelés ci-après éléments graphiques g_i d'un alphabet $G = \{g_1, \ldots, g_I\}$ à I éléments dans la langue prédéterminée, avec $1 \leq i \leq M$, respectivement à des chaînes phonétiques CP composées chacune d'un ou plusieurs éléments phonétiques p_j d'un alphabet $P = \{p_1, \ldots, p_J\}$ à J éléments phonétiques avec $1 \leq j \leq J$ et $I \neq J$ a priori. Toutefois, on ignore à ce stade la segmentation de la chaîne CG en syllabes ou en graphèmes comprenant chacun un ou plusieurs éléments graphiques, et la segmentation de la chaîne CP en phonèmes comprenant chacun un ou plusieurs éléments phonétiques.

Typiquement, les alphabets G et P ont une trentaine d'éléments. Ils présentent ainsi une possibilité de 30 x 30 = 900 couples possibles d'élément graphique et d'élément phonétique. En pratique, le corpus C contient au moins 100.000 transcriptions globales de chaînes typographiques CG en chaînes phonétiques CP, ce qui préserve l'invention d'erreurs grossières dans des estimations de probabilités, comme on le verra ci-après.

A l'étape E2, des premières probabilités de transcription élémentaire $P(g_i|p_j)$ pour qu'un élément

graphique g_i corresponde à l'élément phonétique p_j sont en priorité estimées et enregistrées dans la base de données avec le corpus de transcriptions globales C.

Les valeurs estimées des premières probabilités sont autant que possible proches respectivement de valeurs de probabilité maximales recherchées afin que le procédé de l'invention opérant par itérations converge rapidement tout en évitant de retenir des maxima locaux.

La nature concaténative des transcriptions globales des chaînes conduit à l'hypothèse d'une corrélation entre le rang r_g des éléments graphiques dans une chaîne graphique CG et le rang r_p des éléments phonétiques dans la chaîne phonétique correspondante CP. Par exemple dans la transcription globale (beau|bo), il est plus probable que l'élément graphique b, de par sa position en début de chaîne CG, se traduise en élément phonétique [b] plutôt qu'il ne se traduise en [o] phonétique positionné en fin de la chaîne correspondante CP. Dans cet exemple, la corrélation des rangs rapproche les éléments graphiques [b] et [e] de l'élément phonétique [b], et les éléments graphiques [a] et [u] de l'élément phonétique [o].

L'algorithme d'estimation initiale E2 des premières probabilités $P(g_i|p_j)$ comprend des sous-étapes suivantes E21 à E27.

A la sous-étape E21, IJ nombres de contingence K_{gipj}, respectivement associés aux transcriptions élémentaires $(g_i|p_j)$ d'un élément graphique de l'alphabet G et d'un élément phonétique de l'alphabet P sont mis à zéro. Le nombre de contingence K_{gipj} est égal à la fin de l'étape E2 au nombre de fois estimé où l'élément graphique g_i est retranscrit en

l'élément phonétique p_j dans les diverses transcriptions globales de chaînes typographiques CG en chaînes phonétiques CP incluses dans le corpus C.

Pour chaque transcription de chaîne (CG|CP), comme indiqué à la sous-étape E22, les rangs des éléments graphiques dans la chaîne CG et les rangs des éléments phonétiques dans la chaîne CP sont normalisés en fonction des longueurs respectives l_g et l_p des chaînes CG et CP qui peuvent être différentes. A la sous-étape E23, le rang r d'un élément phonétique dans la chaîne CP est déduit du rang r_{gi} d'un élément graphique g_i dans la chaîne CG auquel sera associé l'élément phonétique de rang r, selon la relation suivante :

$$r = \text{partie entière } (r_{gi}.l_p/l_g).$$

Le nombre de contingences K_{gipj} associé à la transcription élémentaire de l'élément graphique g_i en l'élément phonétique p_j n'est alors incrémenté de 1 que si l'élément phonétique p_j est situé au rang déduit r dans la chaîne CP, comme indiqué aux sous-étapes E24 et E25.

Les sous-étapes E22 à E25 son réitérées pour chaque transcription globale (CG|CP) du corpus C, comme indiqué à la sous-étape E26. Lorsque toutes les transcriptions globales du corpus ont été parcourues, la sous-étape suivante 26 estime toutes les premières probabilités $P(g_i|p_j)$ de transcription élémentaire entre les éléments graphiques et les éléments phonétiques, selon les relations suivantes pour chaque élément graphique g_i :

$$P(g_i|p_j) = K_{gipj} \Big/ \sum_{j=1}^{j=J} Kgipj$$

après avoir calculé le terme somme au dénominateur pour l'élément graphique g_i.

En revenant à la figure 1, le procédé de correspondance est poursuivi par des étapes E3 à E10 qui segmentent chaque chaîne graphique CG lue dans le corpus de la base de données afin de faire correspondre automatiquement d'une manière biunivoque chaque segment de la chaîne CG, appelé graphème, comprenant un ou plusieurs éléments graphiques à un segment, appelé phonème, comprenant un ou plusieurs éléments phonétiques résultant d'une segmentation de la chaîne phonétique correspondante CP.

Une chaîne graphique CG comprend M éléments graphiques consécutifs g_1 à g_M et la chaîne phonétique CP correspondant à la chaîne CG comprend N éléments phonétiques consécutifs p_1 à p_N avec l'entier N différent, ou éventuellement égal à l'entier M.

La probabilité $P(g_1,\ldots g_m,\ldots g_M|p_1,\ldots p_n,\ldots p_N)$ pour que la chaîne CG corresponde à la chaîne CP, avec $1 \leq m \leq M$ et $1 \leq n \leq N$, est déterminée en fonction des premières probabilités de transcription élémentaire $P(g_i|p_j)$ estimées et enregistrées précédemment à l'étape E2, et d'une similarité entre les chaînes CG et CP. La similarité est basée sur la distance d'édition de Damerau-Levenshtein DLM (Damerau-Levenshtein Metric), mais en effectuant une maximalisation et non une minimisation. La probabilité P(CG|CP) est déterminée par une programmation dynamique, en utilisant la formule d'itération suivante pour tout couple m,n tel que $1 \leq n \leq N$ et $1 \leq m \leq M$:

$$P(g_1g_2\ldots g_m|p_1p_2\ldots p_n)=P(g_m|p_n)\max[P(g_1g_2\ldots g_{m-1}|p_1p_2\ldots p_n), P(g_1g_2\ldots g_m|p_1p_2\ldots p_{n-1}), P(g_1g_2\ldots g_{m-1}|p_1p_2\ldots p_{n-1})].$$

La nature concaténative des transcriptions globales de chaînes et des transcriptions graphèmes/phonèmes permet d'appliquer de manière

efficace les modèles de Markov. Pour la probabilité donnée d'une transcription d'une chaîne $g_1,g_2...g_m$ en une chaîne $p_1p_2...p_n$, l'extension de la chaîne graphique, respectivement phonétique, par un nouvel élément graphique g_{m+1}, respectivement p_{n+1}, donne lieu soit à la même chaîne phonétique, respectivement graphique, soit à l'adjonction d'un nouvel élément phonétique, respectivement graphique. Exprimé en terme de probabilité, $P(g_1g_2...g_{m+1}|p_1p_2...p_{n+1})$ ne dépend que des probabilités de trois transcriptions possibles :

soit $P(g_1g_2...g_m|p_1p_2...p_{n+1})$

soit $P(g_1g_2...g_{m+1}|p_1p_2...p_n)$

soit $P(g_1g_2...g_m|p_1p_2...p_n)$.

Cette dépendance est exprimée par la distance d'édition égale à la plus grande des trois probabilités indiquée ci-dessus.

Après avoir mis les indices m et n à zéro pour une transcription globale (CG|CP) à l'étape E3 et incrémenté les indices m et n de 1 aux étapes E4 et E5, des itérations commencent aux étapes E6 et E7 en déterminant les probabilités pour que les M concaténations successives des éléments graphiques g_1 à g_M de la chaîne CG correspondent au premier élément phonétique p_1 de la chaîne CP, soit :

$$P(g_1,...g_m|p_1) = P(g_m|p_1) \max[P(g_1,...g_{m-1}|p_1)]$$

avec $1 \leq m \leq M$, en commençant par la probabilité élémentaire $P(g_1|p_1)$. Puis comme illustré par l'étape E8, le procédé est poursuivi par des itérations pour déterminer les probabilités pour que les M concaténations des éléments graphiques g_1 à g_M de la chaîne CG correspondent aux deux premiers éléments phonétiques p_1 et p_2 de la chaîne CP, en utilisant les probabilités précédemment déterminées pour le premier élément graphique p_1, soit :

$P(g_1,...g_m|p_1, p_2) = P(g_m|p_2) \max[P(g_1,...g_{m-1}|p_2), P(g_1,...g_m|p_1), P(g_1,...g_{m-1}|p_1)]$.

Puis le procédé est poursuivi en ajoutant un élément phonétique p_n pour déterminer les M probabilités $P(g_1|p_1,...p_n)$ à $P(g_1,...,g_M|p_1,...p_n)$ jusqu'aux M probabilités relatives à la chaîne CP = $(p_1,...p_N)$. L'ordinateur construit et mémorise progressivement par itérations des étapes E4 à E8, une matrice de deuxièmes probabilités $P(g_1,...g_m|p_1,...p_n)$ à M colonnes pour concaténations successives des M éléments graphiques et à N lignes pour concaténations successives des N éléments phonétiques, en opérant ligne par ligne selon l'exemple ci-dessus et en commençant par la probabilité $P(g_1|p_1)$ et en finissant par la probabilité $P(g_1,...g_M|p_1,...p_N)$.

Chaque itération relative à la (m.n)ième transcription $[(g_1,...g_m)|(p_1,...p_n)]$ établit un lien entre le couple (g_m,p_n) et le couple à la plus grande probabilité des trois probabilités déterminées précédemment parmi les trois couples (g_{m-1},p_n), (g_m,p_{n-1}) et (g_{m-1},p_{n-1}). Le lien est mémorisé dans l'ordinateur. Lorsque le couple (g_m,p_n) est relié au couple (g_{m-1},p_n), il s'agit d'une transcription élémentaire de (g_{m-1},g_m) en g_m ; lorsque le couple (g_m,p_n) est relié au couple (g_m,p_{n-1}), il s'agit d'une transcription élémentaire de g_m en (p_{n-1},p_n) ; et lorsque le couple (g_m,p_n) est relié au couple (g_{m-1},p_{n-1}), il s'agit d'une transcription élémentaire de g_m en p_n.

Ainsi à chaque détermination de probabilité $P(g_1,...g_m)|(p_1,...p_n)$ est mémorisé dans l'ordinateur un lien. Les liens tracent un chemin unique également mémorisé progressivement dans l'ordinateur et reliant le premier couple (g_1, p_1) au dernier couple (g_M, p_N)

dans la matrice à M colonnes et N lignes. La topologie du chemin unique dans la matrice de taille M.N segmente les chaînes graphiques CG en graphèmes et les chaînes phonétiques CP en phonèmes et aligne les éléments graphiques et les éléments phonétiques en correspondance biunivoque. Si un segment du chemin suit une portion d'une ligne entre deux éléments graphiques, la concaténation des éléments graphiques de la portion de ligne correspond à l'élément phonétique de la ligne complété par un ou des éléments phonétiques muets et sans signification afin de former un couple de graphème et de phonème ayant le même nombre d'éléments, lequel couple est mémorisé dans l'ordinateur. Si un segment du chemin suit une portion de colonne entre deux éléments phonétiques, l'élément graphique de la colonne complété par un ou des éléments graphiques sans signification correspond à la concaténation des éléments phonétiques de la portion de colonne afin de former un couple de graphème et de phonème ayant le même nombre d'éléments, lequel couple est mémorisé dans l'ordinateur. Un changement de direction du chemin vers l'horizontale, la verticale ou la diagonale dans la matrice indique une segmentation des chaînes CG et CP.

A titre d'exemple simple, on cherche à segmenter la transcription globale du mot CG = "beau" en la chaîne phonétique CP = [bo] en supposant que l'étape E2 a estimé les premières probabilités individuelles suivantes dans le corpus C :

$P(b|b)=0,9$; $P(e|b)=0,1$; $P(a|b)=0,1$; $P(u|b)=0,1$

$P(e|o)=0,2$; $P(a|o)=0,1$; $P(u|o)=0,2$; $P(b|o)=0,1$.

Pour la transcription (beau|bo) du corpus, les M=4 itérations des étapes E5, E6 et E7 pour chacune

des M=2 lignes de la matrice de taille (4,2) produisent le tableau suivant:

p_n / g_m	b = g_1	e = g_2	a = g_3	u = g_4
[b] = p_1	0,9	←0,09	←0,09	←0,0009
[o] = p_2	↑0,09	↖0,18	←0,018	←0,0036

Le symbole ← indique que le couple (g_m , p_n) est relié au couple (g_{m-1}, p_n); le symbole ↑ indique que le couple (g_m , p_n) est relié au couple (g_m, p_{n-1}) ; et le symbole ↖ indique que le couple (g_m , p_n) est relié au couple (g_{m-1}, p_{n-1}). Le symbole ↖ associé à la transcription (be|bo) indique que cette dernière est déduite et donc liée à la transcription (b|b) qui la précède. Le symbole ↖ indique une frontière de segmentation entre des couples de graphème et phonème. On en déduit de ce tableau l'alignement suivant :

b eau
b o**.

Le symbole * désigne un élément phonétique muet et sans signification.

Afin de parfaire les correspondances entre les graphèmes et les phonèmes et les correspondances entre les éléments graphiques et les éléments phonétiques, de préférence comme indiqué par l'étape E11, les premières probabilités $P(g_1|p_1)$ à $(P(g_I|p_J)$ des transcriptions de chacun des éléments graphiques respectivement en les J éléments phonétiques (étape E2) et en particulier les nombres de contingence K_{g1p1} à K_{gIpJ} (sous-étape E25) sont à nouveau estimés en fonction notamment des rangs des éléments phonétiques placés dans les chaînes phonétiques données CG qui ont été segmentées en phonèmes à l'étape précédente E10. A nouveau des deuxièmes

probabilités $P(g_1,...g_m|p_1,...p_n)$ de MN deuxièmes transcriptions de chaque transcription globale d'une chaîne graphique donnée à M éléments graphiques (CG) en une chaîne phonétique correspondante (CP) à N éléments phonétiques sont déterminées par l'exécution des étapes E3 à E10 afin qu'à l'étape suivante E10 des liens soient établis entre des couples (g_m,p_n) d'une nouvelle matrice à M colonnes et N lignes et par conséquent un chemin corrigé reliant le dernier couple (g_M,p_N) au premier couple (g_1,p_1) dans la nouvelle matrice de deuxièmes probabilités de taille MN.

Eventuellement, grâce à la capacité et la rapidité élevées de traitement de l'ordinateur, d'autres boucles itératives d'étapes E2 à E11 peuvent être exécutées dans l'ordinateur jusqu'à la convergence du procédé de correspondance, c'est-à-dire jusqu'à ce que le chemin établi devienne constant d'une boucle à la suivante.

Après la segmentation de toutes les chaînes graphiques et phonétiques du corpus G en graphèmes et phonèmes, la base a enregistrée toutes les correspondances entre les éléments graphiques et phonétiques et les correspondances entre les graphèmes et phonèmes pour tout le corpus C parcouru.

Toute nouvelle chaîne graphique ajoutée au corpus peut être ensuite automatiquement transcrite en une chaîne phonétique segmentée en des phonèmes à l'aide notamment des correspondances précédemment établies et enregistrées selon l'invention, ce qui enrichit progressivement le corpus dans la base de données et augmente la précision des transcriptions.

Comme déjà dit, les transcriptions phonétiques sont utiles à des systèmes logiciels correcteurs

orthographiques de fautes pour reconnaître des fautes lexicales lors de la saisie de texte sur un clavier de terminal. Ainsi lorsque la nouvelle chaîne graphique ajoutée au corpus est saisie sur un clavier d'un terminal, la chaîne phonétique segmentée en phonèmes au moyen des correspondances enregistrées est utilisée pour une correction orthographique de la nouvelle chaîne graphique saisie.

Le procédé de l'invention peut être également utilisé comme outil de génération automatique de messages courts SMS à partir d'un texte rédigé dans la langue courante. Il nécessite pour ce faire un corpus d'apprentissage C dont les transcriptions sont adaptées à la génération automatique de messages courts et font correspondre respectivement des chaînes graphiques CG, telles que des mot et des locutions, à des chaînes phonétiques CP dont les "phonèmes" sont phonétiquement lisibles par toute personne non phonéticienne. Par exemple, le corpus établit les correspondances en français suivantes entre chaînes graphiques et chaînes phonétiques:

j'ai : G
air : R
occupé : OQP
cas : K.

Ainsi une nouvelle chaîne graphique saisie dans un terminal est automatiquement transcrite par le procédé de l'invention en une chaîne phonétique segmentée en phonèmes lisibles par toute personne non phonéticienne au moyen des correspondances enregistrées pour être incluse dans un message court. Selon l'exemple précédent, la phrase en français "j'ai l'air occupé" saisie dans le terminal est transcrite automatiquement en un message court de suivant Gl'ROQP à transmettre par le terminal, les

"chaînes phonétiques" [G], [l'], [R] et [OQP] étant phonétiquement lisibles par tout usager non phonéticien. En variante, les "chaînes phonétiques" [G], [l'], [R] et [OQP] peuvent être assimilées à des éléments phonétiques pour constituer une chaîne phonétique [Gl'ROQP].

Selon une implémentation préférée du procédé de l'invention, les étapes du procédé de l'invention sont déterminées par les instructions d'un programme d'ordinateur incorporé dans un ordinateur tel qu'un terminal, un ordinateur personnel, un serveur ou tout autre système informatique. Le programme fait correspondre automatiquement des éléments graphiques composant des chaînes graphiques données à des éléments phonétiques composant des chaînes phonétiques correspondantes, après avoir saisi initialement des transcriptions globales des chaînes graphiques en les chaînes phonétiques dans une base accessible par l'ordinateur et avoir estimé et enregistré dans la base des premières probabilités de transcriptions élémentaires des éléments graphiques respectivement en les éléments phonétiques. Le programme comporte des instructions de programme qui, lorsque ledit programme est chargé et exécuté dans l'ordinateur dont le fonctionnement est alors commandé par l'exécution du programme, réalisent les étapes du procédé selon l'invention.

En conséquence, l'invention s'applique également à un programme d'ordinateur, notamment un programme d'ordinateur sur ou dans un support d'informations, adapté à mettre en œuvre l'invention. Ce programme peut utiliser n'importe quel langage de programmation, et être sous la forme de code source, code objet, ou de code intermédiaire entre code

source et code objet tel que dans une forme partiellement compilée, ou dans n'importe quelle autre forme souhaitable pour implémenter le procédé selon l'invention.

REVENDICATIONS

1 - Procédé mis en oeuvre dans un ordinateur pour faire correspondre automatiquement des éléments graphiques (g_i) composant des chaînes graphiques données à des éléments phonétiques (p_j) composant des chaînes phonétiques correspondantes, après avoir saisi (E1) initialement des transcriptions globales (CG|CP) des chaînes graphiques en les chaînes phonétiques dans une base accessible par l'ordinateur et avoir estimé et enregistré dans la base (E2) des premières probabilités ($P(g_i|p_j)$) de transcriptions élémentaires des éléments graphiques respectivement en les éléments phonétiques, caractérisé par les étapes suivantes :

pour chaque transcription d'une chaîne graphique donnée (CG) à M éléments graphiques en une chaîne phonétique correspondante (CP) à N éléments phonétiques, déterminer (E3 - E9) des deuxièmes probabilités ($P(g_1,...g_m|p_1,...p_n)$) de MN deuxièmes transcriptions de M chaînes graphiques concaténant successivement les M éléments graphiques en N chaînes phonétiques concaténant successivement les N éléments phonétiques, en fonction chacune d'une première probabilité respective et de la plus grande de trois deuxièmes probabilités respectives déterminées précédemment, et

établir et mémoriser (E10) un lien entre les derniers éléments (g_m,p_n) des chaînes graphique et phonétique de chaque deuxième transcription et les derniers éléments des chaînes graphique et phonétique de la transcription relative à la plus grande des trois deuxièmes probabilités respectives afin que des liens établis dans une matrice de taille MN relative aux deuxièmes probabilités constitue un chemin unique

entre des dernier et premier couples d'éléments graphique et phonétique de la matrice pour segmenter la chaîne graphique donnée en des graphèmes correspondant respectivement à des phonèmes segmentant la chaîne phonétique correspondante et pour enregistrer les correspondances entre les graphèmes et phonèmes dans la base, le nombre d'éléments graphiques dans un graphème étant identique au nombre d'éléments phonétiques dans le phonème correspondant, afin que toute nouvelle chaîne graphique soit automatiquement transcrite en une chaîne phonétique segmentée en phonèmes au moyen des correspondances enregistrées.

2 - Procédé conforme à la revendication 1, selon lequel la première probabilité respective pour la détermination (E3 - E9) d'une deuxième probabilité ($P(g_1,...g_m|p_1,...p_n)$) relative à une deuxième transcription d'une chaîne graphique concaténant m éléments graphiques en une chaîne phonétique concaténant n éléments phonétiques, avec $1 \leq m \leq M$ et $1 \leq n \leq N$, est relative aux derniers éléments dans la chaîne graphique à m éléments graphiques et la chaîne phonétique à n éléments phonétiques.

3 - Procédé conforme à la revendication 1 ou 2, selon lequel les trois deuxièmes probabilités respectives déterminées précédemment pour la deuxième transcription de la chaîne graphique à m éléments graphiques en la chaîne phonétique à n éléments phonétiques sont respectivement relatives à une deuxième transcription d'une chaîne graphique à m-1 éléments graphiques en la chaîne phonétique à n éléments phonétiques, une deuxième transcription de la chaîne graphique à m éléments graphiques en une

chaîne phonétique à n-1 éléments phonétiques et une deuxième transcription de la chaîne graphique à m-1 éléments graphiques en la chaîne phonétique à n-1 éléments phonétiques.

4 - Procédé conforme à l'une quelconque des revendications 1 à 3, comprenant une estimation d'autres premières probabilités ($P(g_i|p_j)$) de transcriptions de chacun des éléments graphiques respectivement en les éléments phonétiques en fonction notamment des rangs des éléments phonétiques placés dans les chaînes phonétiques données (CG) qui ont été segmentées en phonèmes afin à nouveau de déterminer (E6) des deuxièmes probabilités ($P(g_1,...g_m|p_1,...p_n)$) de MN deuxièmes transcriptions de chaque transcription d'une chaîne graphique donnée à M éléments graphiques (CG) en une chaîne phonétique correspondante (CP) à N éléments phonétiques et établir un chemin corrigé reliant le dernier couple (g_M,p_N) au premier couple (g_1,p_1) dans une nouvelle matrice de deuxièmes probabilités de taille MN.

5 - Procédé conforme à l'une quelconque des revendications 1 à 4, selon lequel la nouvelle chaîne graphique est saisie sur un clavier d'un terminal et la chaîne phonétique segmentée en phonèmes au moyen des correspondances enregistrées est utilisée pour une correction orthographique de la nouvelle chaîne graphique saisie.

6 - Procédé conforme à l'une quelconque des revendications 1 à 4, selon lequel les chaînes phonétiques sont phonétiquement lisibles par toute personne non phonéticienne, et la nouvelle chaîne graphique est automatiquement transcrite en une

chaîne phonétique segmentée en phonèmes lisibles par toute personne non phonéticienne au moyen des correspondances enregistrées pour être incluse dans un message court.

7 - Programme d'ordinateur apte à être mis en oeuvre dans un ordinateur pour faire correspondre automatiquement des éléments graphiques (g_i) composant des chaînes graphiques données à des éléments phonétiques (p_j) composant des chaînes phonétiques correspondantes, après avoir saisi (E1) initialement des transcriptions globales (CG|CP) des chaînes graphiques en les chaînes phonétiques dans une base accessible par l'ordinateur et avoir estimé et enregistré dans la base (E2) des premières probabilités ($P(g_i|p_j)$) de transcriptions élémentaires des éléments graphiques respectivement en les éléments phonétiques, ledit programme comprenant des instructions qui, lorsque le programme est chargé et exécuté dans l'ordinateur, réalisent les étapes suivantes:

pour chaque transcription d'une chaîne graphique donnée (CG) à M éléments graphiques en une chaîne phonétique correspondante (CP) à N éléments phonétiques, déterminer (E3 - E9) des deuxièmes probabilités ($P(g_1,...g_m|p_1,...p_n)$) de MN deuxièmes transcriptions de M chaînes graphiques concaténant successivement les M éléments graphiques en N chaînes phonétiques concaténant successivement les N éléments phonétiques, en fonction chacune d'une première probabilité respective et de la plus grande de trois deuxièmes probabilités respectives déterminées précédemment, et

établir et mémoriser (E10) un lien entre les derniers éléments (g_m,p_n) des chaînes graphique et

phonétique de chaque deuxième transcription et les derniers éléments des chaînes graphique et phonétique de la transcription relative à la plus grande des trois deuxièmes probabilités respectives afin que des liens établis dans une matrice de taille MN relative aux deuxièmes probabilités constitue un chemin unique entre des dernier et premier couples d'éléments graphique et phonétique de la matrice pour segmenter la chaîne graphique donnée en des graphèmes correspondant respectivement à des phonèmes segmentant la chaîne phonétique correspondante et pour enregistrer les correspondances entre les graphèmes et phonèmes dans la base, le nombre d'éléments graphiques dans un graphème étant identique au nombre d'éléments phonétiques dans le phonème correspondant, afin que toute nouvelle chaîne graphique soit automatiquement transcrite en une chaîne phonétique segmentée en phonèmes au moyen des correspondances enregistrées.

FIG. 1

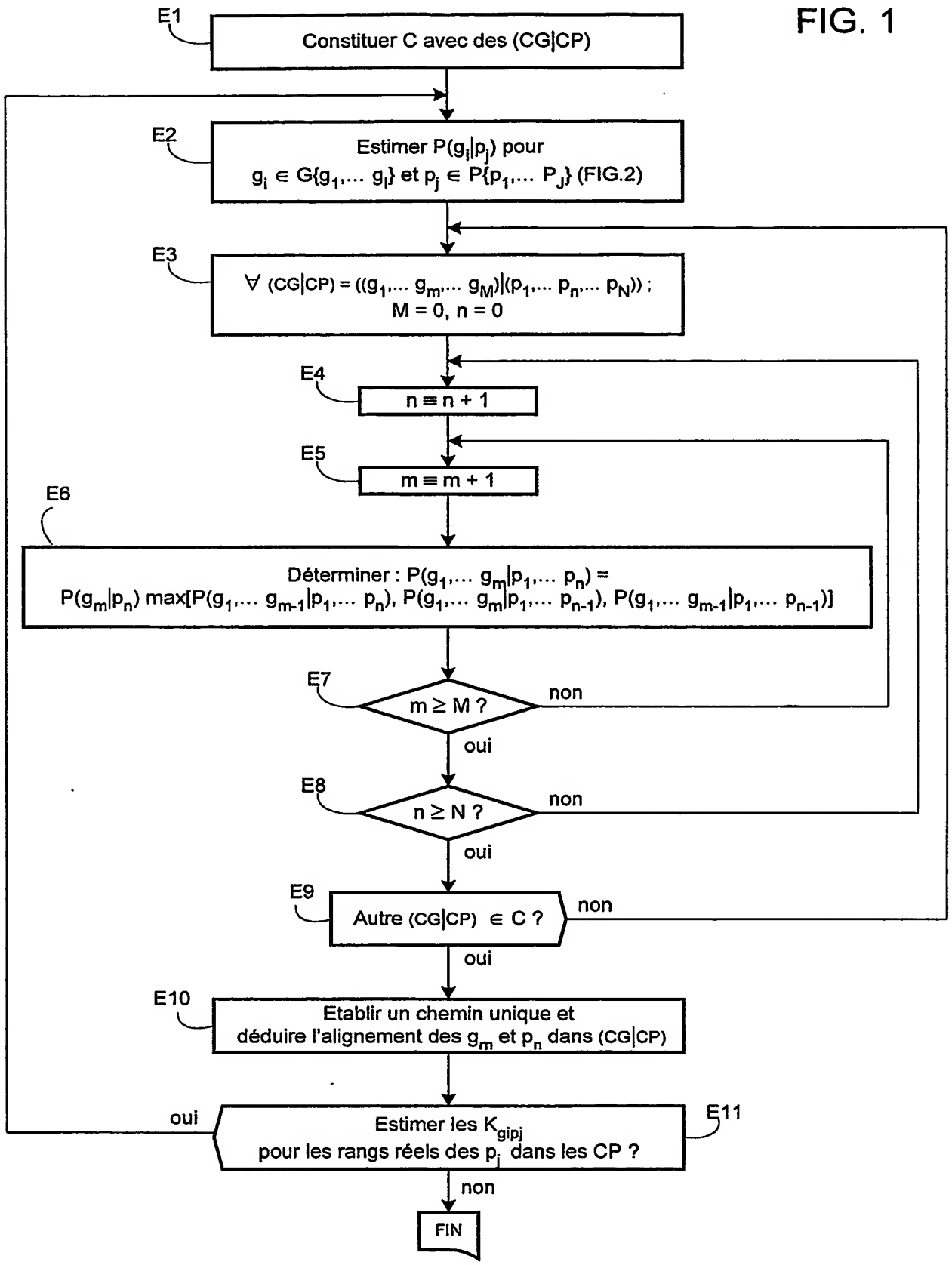
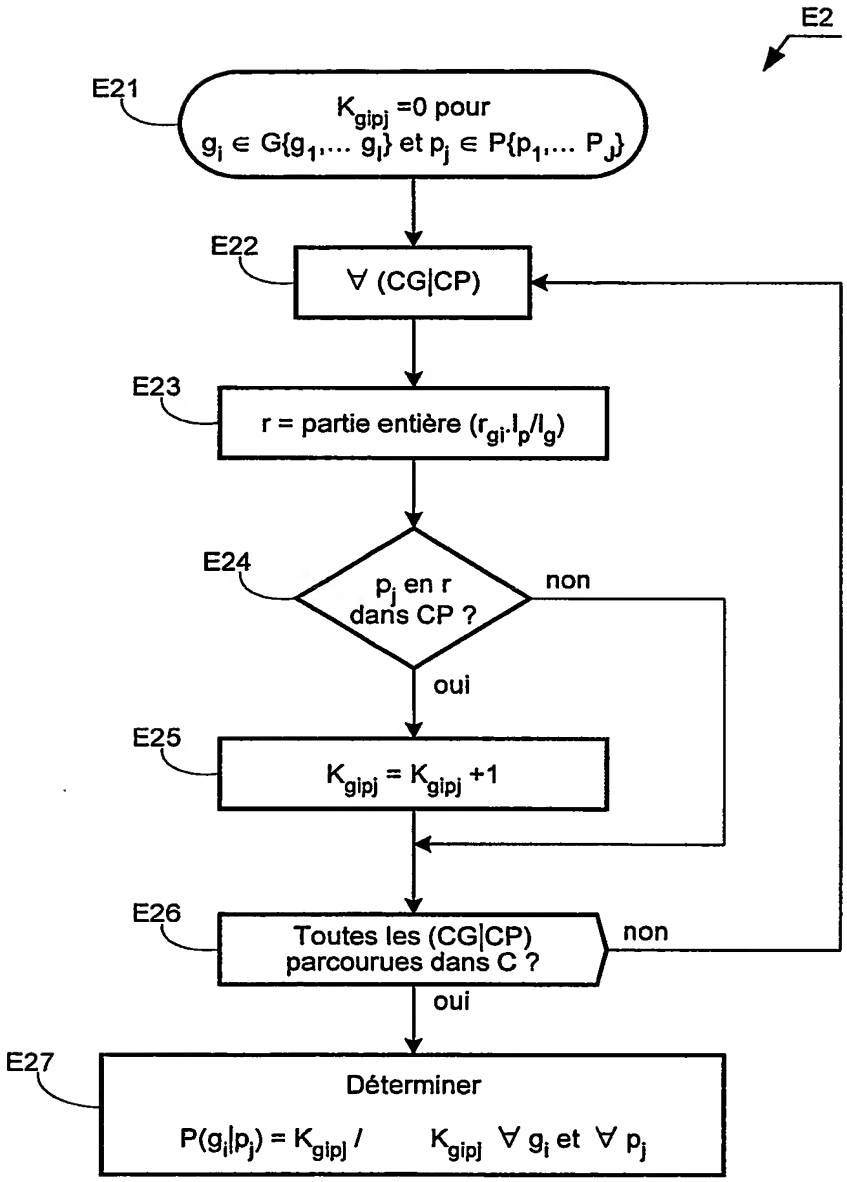

E1
Constituer C avec des (CG|CP)
E2
Estimer P(g_i|p_j) pour
g_i ∈ G{g_1,... g_I} et p_j ∈ P{p_1,... P_J} (FIG.2)
E3
∀ (CG|CP) = ((g_1,... g_m,... g_M)|(p_1,... p_n,... p_N)) ;
M = 0, n = 0
E4
n ≡ n + 1
E5
m ≡ m + 1
E6
Déterminer : P(g_1,... g_m|p_1,... p_n) =
P(g_m|p_n) max[P(g_1,... g_m-1|p_1,... p_n), P(g_1,... g_m|p_1,... p_n-1), P(g_1,... g_m-1|p_1,... p_n-1)]
E7
m ≥ M ?
non
oui
E8
n ≥ N ?
non
oui
E9
Autre (CG|CP) ∈ C ?
non
oui
E10
Etablir un chemin unique et
déduire l'alignement des g_m et p_n dans (CG|CP)
oui
Estimer les K_gipj
pour les rangs réels des p_j dans les CP ?
E11
non
FIN

FIG. 2


E2
E21
K_{gipj} =0 pour
$g_i \in G\{g_1,\ldots g_I\}$ et $p_j \in P\{p_1,\ldots P_J\}$
E22
∀ (CG|CP)
E23
r = partie entière ($r_{gi}.I_p/I_g$)
E24
p_j en r
dans CP ?
non
oui
E25
$K_{gipj} = K_{gipj} +1$
E26
Toutes les (CG|CP)
parcourues dans C ?
non
oui
E27
Déterminer
$P(g_i|p_j) = K_{gipj} /$ K_{gipj} ∀ g_i et ∀ p_j